I put the issues re: Urdu to the Bidi committee and got this feedback:

From: "Mansour, Kamal" <kamal.mansour@agfamonotype.com>
Subject: RE: Proposal for Urdu (L2/02-163)
Date: Wed, 1 May 2002 19:45:29 -0500

On Wed, 01 May 2002 14:40, Asmus Freytag wrote:

> **1. Misra sign** - this seems a unique character and acceptable
> for encoding

The documentation shows reasonable evidence for this unique character.
It certainly fits the pattern of 'para-alphabetic' symbols used in Urdu.

> **4. Jazm Urdu** - there was considerable discussion on whether the
> distinctions between this and Sikkun(?) needs to be made in
> plain text.

I think it would be a mistake to encode Urdu Jazm because it is a
locale-specific variant of Sukun.

> **5. Small high tah** - this seems a unique character and acceptable for
> encoding, can someone verify this against Unicode 3.2?

I think we need more contextual information on this character. Is its
behavior similar to the Koranic annotation signs such as 06DA or 06E2?

> **6. Bismallah** - the document makes an interesting argument for
> why this should be encoded. There's a sentiment to encode this.

This graphic for this phrase (Bismillah) is used with such high frequency,
it has become almost iconic in nature. I'm surprised it hasn't been
proposed before.

Kamal

Jonathan Kew also responded:

Date: Thu, 02 May 2002 11:31:18 +0100
Subject: Re: [bidi] Proposal for Urdu (L2/02-163)
From: Jonathan Kew <jonathan_kew@sil.org>
On 1/5/02 10:39 pm, Asmus Freytag at asmusf@ix.netcom.com wrote:

> UTC is just discussing this document (also available as WG2/N2413). There
> are various issues raised that would be nice to get input from people on
> this list with expertise in Urdu and Arabic.
>
> **1. Misra sign** - this seems a unique character and acceptable for encoding

That seems fine.

> **2. Sign Safah** - this may be the same as the Arabic number sign we've
> already approved (see document L2/02-106)

No, it is not. The form is different (derived from the letter SAD, whereas the number sign generally looks like it begins with an AIN or HAMZA-like shape, though I suspect its origin may be a vestigial "skeleton" of the Urdu spelling of "number"!), and the meaning is distinct: "Page" as opposed to "Number".

So I would support the separate encoding of this sign. Note that it should be of the same category as the number, year, and footnote signs recently accepted, so that it is considered to "apply" to the following sequence of digits (exact scope more carefully spelled out elsewhere).

> **3. Nuqtatain** - is there confirmation that this is really needed as
> composite and that its appearance is different from 06D4 combined with a
> colon in an Arabic font. Is this not sufficient to code a different from of
> Hyphen so it can be composed.

I'm doubtful about this one. It seems to me, both visually and semantically, to be a composite of COLON plus HYPHEN, with the glyphs stylized appropriately for the font. (One sometimes sees the exact same thing in English text, more often hand-written than printed:- that's probably where Urdu picked it up from.) You could even have a colon-hyphen ligature in a font, given a "smart" rendering system, if a particular appearance is desired that differs from the components by themselves.

U+06D4 is ARABIC FULL STOP (correct name?), which is not the appropriate character to use here. The form shown in the code charts is derived from an elongated "flick" of the pen typically used in Nastaliq calligraphy; in other type styles, a full stop may well appear as a round dot. As such, I don't think this should be used in composing the "nuqtatain". Also, I have often seen COLON + HYPHEN (with a simple "straight" appearance) used for this function in non-Nastaliq Arabic-script text.

> **4. Jazm Urdu** - there was considerable discussion on whether the
> distinctions between this and Sikkun(?) needs to be made in plain text.

While I would have no real objection to this, I don't think there is a strong case. Logically, it is the same character as SUKUN; the different shape reflects calligraphic tradition, not character identity. Moreover, we already have a second "SUKUN/JAZM" encoded, though mis-named, at U+06E1, so it is possible to distinguish two SUKUNs in plain text using this.

It has been pointed out that there is a *third* distinct SUKUN-like character in the King Fahd Qur'an, and this probably merits encoding, but I don't think it has been formally proposed as yet.

> **5. Small high tah** - this seems a unique character and acceptable for
> encoding, can someone verify this against Unicode 3.2?

It's not currently encoded, and looks acceptable from the example given.

However, as something very similar-looking occurs as a component of retroflex Urdu letters such as U+0688, we risk accidentally starting down the road of supporting dynamic composition of new Arabic-script letters. Will it be valid for someone to create a PUNJABI RETROFLEXED NOON (which has a dot and a small TAH, unlike the Sindhi version already encoded as U+06BB; expect a proposal shortly!) by using U+0646 with this new mark? What about the use of this sign to create retroflexed vowels, by applying it to ALEF, YEH, WAW (this has been proposed for at least one language in northern Pakistan)?

Do we endorse this kind of usage, and if so what should the combining class and positioning behavior be for this mark (which in the example shown from Arabic and Urdu looks more akin to the Qur'anic marks)? And if not, how do we ensure that users are not tempted to abuse the mark in this way?

> **6. Bismallah** - the document makes an interesting argument for why this
> should be encoded. There's a sentiment to encode this.

No comment, except that I'm a bit surprised.

> **7. Digits** - the arguments presented aren't considered sufficient to change
> the unification of Farsi and Urdu digits and UTC formally rejected these
> and intends to better document the status quo.

OK... but something I've been wondering: would there be a case for defining variant selectors for the digits where Urdu, Sindhi, or other languages have distinct forms? In the case of something like the FOUR or SEVEN, it seems to me that the Urdu forms are substantively different from the Arabic or Farsi ones, not merely stylistic variants, and it therefore seems reasonable for users to be able to encode "URDU DIGIT FOUR", as distinct from "FARSI DIGIT FOUR", explicitly in plain text. If the UTC is opposed to encoding the Urdu digits in their own right, variant selectors would be another way to support this.

Jonathan

To which Paul Nelson added:

```
Date: Thu, 2 May 2002 03:48:34 -0700
From: "Paul Nelson" <paulnel@microsoft.com>

I personally believe that the correct way to get the shapes is to make
this a formatting option and allow smart fonts to correctly render the
shape. Symantically there is no difference between showing a Farsi "4"
vs. an Urdu "4" vs. a Latin "4". The digit value is still understood to
be the same. In the plain text form the digit value is understood by
users, even if they don't like the shape...and they can always change
the font. To add some Variant character into the data stream will make
parsing text more time consuming and will end up requiring smart font
technology to be used similar to how I can already achieve displaying
the correct digit shapes more efficiently. I fear that with the use of
variant characters in this manner we begin to turn Unicode into a
formatting language as well.
```